Topic Detection & Tracking (TDT)

Overview & Perspective

Charles L. Wayne

National Security Agency
Ft. Meade, MD 20755

ABSTRACT

Topic Detection and Tracking (TDT) refers to automatic techniques for finding topically related material in streams of data (e.g., newswire and broadcast news). Work on TDT began about a year ago, is now expanding, and will be a regular feature at future Broadcast News workshops.

INTRODUCTION

Thanks to DARPA and NSF funding for research on critical speech and text processing problems plus NIST sponsorship of objective performance evaluations, the research community has been very successful in recent years in creating critical new component technologies. These include, for example, increasingly accurate and robust techniques for converting speech to text in Hub 4, for finding documents that match *ad hoc* queries in TREC, and for finding named entities in MUC. These technologies enable a wide variety of important applications, and their successes permit new research challenges to be undertaken.

Topic Detection and Tracking (TDT) is an important, relatively new challenge. Work on TDT began about a year ago, is now expanding, and will be a regular feature at future Broadcast News workshops. DARPA is serious about solving this problem and welcomes more sites to participate in the annual evaluations.

DEFINITION & MOTIVATION

TDT refers to automatic techniques for finding topically related material in streams of data — techniques that could be quite valuable in a wide variety of applications where efficient and timely information access is important.

For example, a lot of useful information could be gleaned from a multitude of news sources, but no one has the time to watch, listen to, or read carefully each of the many news sources available.

Figure 1 illustrates the problem and suggests the solution:

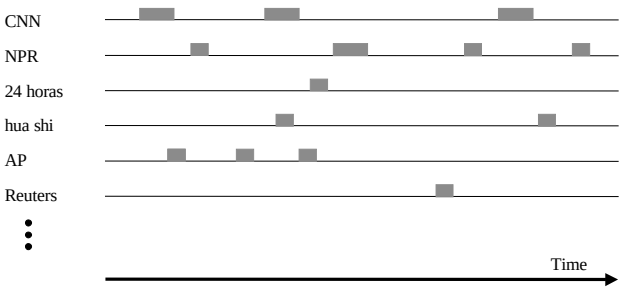

Figure 1: Topic Detection & Tracking in News
(blocks are stories about the same event in several media)

It would be very helpful to have a machine able to map out the data automatically — finding story boundaries, determining what stories go with one another, and discovering when something new (unforeseen) has happened. These capabilities are all encompassed by TDT.

Related problems, not encompassed in the above definition, include estimating the importance of an event (which depends on a user's interests), characterizing an event, determining the redundant information in a set of stories, or summarizing the contents of one or more stories. These are not part of TDT. In addition, TDT has not yet addressed the problem of finding thematically related stories (e.g., about bombings in general); rather, it has focused on finding individual events (e.g., particular bombings).

To solve the TDT challenges, we are looking for robust, accurate, fully automatic algorithms that are source, medium, domain, and language independent.

PILOT STUDY

During 1997 a pilot study was conducted to explore various approaches and to establish performance baselines. Research objectives included finding topically homogeneous regions (*Segmentation*), detecting the occurrence of new events (*Detection*), and tracking the recurrence of known events (*Tracking*), as illustrated in Figure 2:

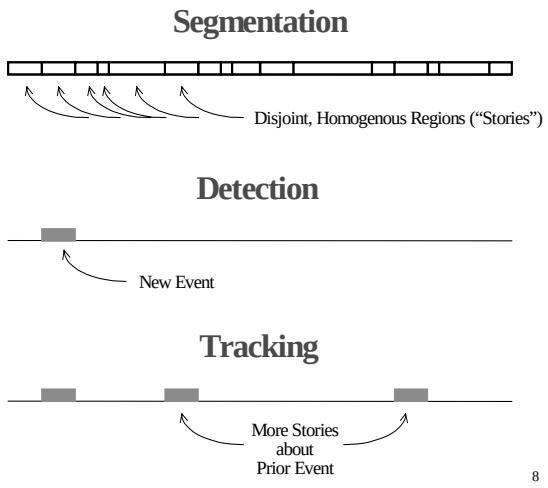

Figure 2: Tasks for Pilot Study

In a highly cooperative team effort, a corpus was collected and annotated, evaluation procedures were devised, software was created, and novel research was conducted by Carnegie Mellon University, Dragon Systems, and the University of Massachusetts. Technical approaches and results were presented at the Topic Detection and Tracking Workshop held at the University of Maryland on 27-28 October 1997. The approaches and results are summarized in the next three papers (by Jon Yamron, Jaime Carbonell and Yiming Yang, and James Allan). A final report will soon be available at http://ciir.cs.umass.edu/projects/tdt.

Corpus

The TDT Pilot Study Corpus consists of 15,863 news stories from July 1994 through June 1995, about half from newswire (Reuters) and half from broadcast news (CNN), the latter in the form of manual transcriptions (no audio).

To help focus the pilot study, the definition of "topic" was narrowed to "event", and researchers chose the 25 events listed in Table 1 as target events.

1	Aldrich Ames
2	Carlos the Jackal
3	Carter in Bosnia
4	Cessna on White House
5	Clinic Murders (Salvi)
6	Comet into Jupiter
7	Cuban Riot in Panama
8	Death of Kim Jong Il (N. Korea)
9	DNA in OJ Trial
10	Haiti Ousts Observers
11	Hall's Copter (N. Korea)
12	Humble, TX, Flooding
13	Justice-to-be Breyer
14	Karrigan/Harding
15	Kobe Japan Quake
16	Lost in Iraq
17	NYC Subway Bombing
18	Oklahoma City Bombing
19	Pentium Chip Flaw
20	Quayle Lung Clot
21	Serbians Down F-16
22	Serbs Violate Bihac
23	Shannon Faulkner
24	USAir 427 Crash
25	World Trade Center Bombing

Table 1: Target Events for Pilot Study

For each of these 25 target events, annotators labeled each story in the corpus with YES, NO, or BRIEF (the latter label meaning that the story contained only a passing reference to the particular event). These annotations provided the ground truth for the experiments.

A fuller description of the corpus (including particular sources, annotator instructions, number of stories per event, and SGML format) resides at http://www.ldc.upenn.edu/TDT/Pilot/TDT.Study.Corpus.v1.3.ps. The corpus itself can be obtained from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

Evaluation Methodology

A crisp, clean evaluation methodology was developed for each of the pilot study tasks. Segmentation performance was measured in terms of whether a story boundary was correctly found between two points 250 words apart. (Segmentation performance was also measured indirectly in terms of its influence on tracking.) Detection performance was measured both retrospectively, after processing the entire corpus, and on-line, after processing each story. Tracking performance was measured for 1, 2, 4, 8, and 16 training examples.

Wherever appropriate, Detection Error Tradeoff (DET) curves [1], were used to compare results obtained with different methods or different parameter settings.

A copy of the evaluation specification will be found at http://trec.nist.gov/tdt.html.

ENLARGED STUDY

In 1998, the TDT effort is being enlarged and extended to encompass several new challenges. It will address the same three tasks (Segmentation, Detection, and Tracking) as the pilot study did, but the evaluation procedures will be modified somewhat. In addition, the volume and variety of data and the number of target topics will all be expanded.

Most significantly, the enlarged TDT effort will seriously attack the problems introduced by imperfect, machine-generated transcripts of audio data.

BBN Technologies, Carnegie Mellon University, Dragon Systems, IBM, SRI, the University of Massachusetts, and the University of Pennsylvania all intend to participate in the 1998 TDT evaluations.

Corpus

To support the enlarged TDT effort, the LDC is now collecting a new corpus (hereinafter "TDT2 corpus"). It will contain approximately 40,000 stories and include approximately 1,000 hours of audio material, all from January to June 1998, subdivided into three subcorpora — two months of training data, two months of development test data, two months of evaluation test data.

The TDT2 corpus will contain data from two newswires (*AP WorldStream* plus *New York Times Newservice*), two radio programs (*PRI The World* plus *VOA World News*), and two television programs (*CNN Headline News* plus *ABC World News Tonight*).

Approximately 100 target topics (defined somewhat more broadly than the events used in the pilot study) will be chosen by random selection of stories from the corpus. For each of these topics, each story in the corpus will then be labeled with YES, NO, or BRIEF.

Dragon Systems will produce a baseline set of automatic transcripts for the audio portion of the corpus. These will be distributed in SGML format along with the closed captions that exist for the television programs and manually-generated transcripts (of closed caption quality) for the radio programs.

Plans for collecting and annotating the corpus are explained more fully in the paper by Mark Liberman; plans for automatically transcribing the corpus are outlined in the paper by Michael Newman. The TDT2 corpus will be released in phases to sites participating in the 1998 DARPA TDT evaluation. In 1999, it will be available to everyone through the LDC.

Evaluation Methodology

An initial evaluation plan for TDT2 research, drafted by George Doddington, appears later in this workshop proceedings. Sites that elect to participate in this year's evaluation may help refine that plan. A dry run will be held sometime this summer, and the formal evaluation will be conducted in the winter, after the Hub 4 results are submitted and before the Broadcast News workshop, where both Hub 4 and TDT results will be presented and discussed.

CONCLUSIONS

TDT will be an important DARPA research focus for several years. It presents new and interesting challenges and builds off prior and on-going successes in automatic speech-to-text conversion and information retrieval. Groups that wish to participate in the 1998 evaluation should contact Charles Wayne (clwayne@snap.org).

As research on TDT progresses and starts dealing seriously with audio sources, it will not only take advantage of the great progress that the speech and text communities have already made, it will also motivate new work on critical problems: In speech, prosody may become more valuable; and lowering the average word error rate may not be as important as dealing with new words. In text, especially machine-transcribed text, it may be important to identify and to extract names. (The speech track in TREC will have similar beneficial effects, pulling on existing technology.) Starting in 1999, Spanish and Mandarin data will be added to the TDT challenges.

REFERENCES

1. Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M.; "The DET Curve in Assessment of Detection Task Performance"; *Proceedings of EuroSpeech '97*, Volume 4, pages 1895-1898; European Speech Communication Association (ESCA).